D98–D108 *Nucleic Acids Research, 2012, Vol. 40, Database issue* *Published online 23 November 2011*
doi:10.1093/nar/gkr1032

GeneDB—an annotation database for pathogens

Flora J. Logan-Klumpler[1,10,*], Nishadi De Silva[1], Ulrike Boehme[1], Matthew B. Rogers[1,2], Giles Velarde[1,2,10], Jacqueline A. McQuillan[1], Tim Carver[1], Martin Aslett[1], Christian Olsen[4], Sandhya Subramanian[4], Isabelle Phan[4], Carol Farris[4,5], Siddhartha Mitra[4], Gowthaman Ramasamy[4], Haiming Wang[6], Adrian Tivey[1], Andrew Jackson[1], Robin Houston[1], Julian Parkhill[1], Matthew Holden[1], Omar S. Harb[7], Brian P. Brunk[7], Peter J. Myler[4,5,8], David Roos[9], Mark Carrington[10], Deborah F. Smith[2], Christiane Hertz-Fowler[3] and Matthew Berriman[1,*]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, [2]Centre for Immunology and Infection, Department of Biology, University of York, Heslington, York YO10 5YW, [3]Centre for Genomic Research, Department of Functional and Comparative Genomics, University of Liverpool, Liverpool L69 3BX, UK, [4]Seattle Biomedical Research Institute, Seattle, WA 98109, [5]Department of Medical Education and Biomedical Informatics, University of Washington, Seattle, WA 98195, [6]Center for Tropical and Emerging Global Diseases, University of Georgia, Athens, GA 30602, [7]Penn Center for Bioinformatics, University of Pennsylvania, Philadelphia, PA 19104, [8]Department of Global Health, University of Washington, Seattle, WA 98195, [9]Department of Biology, University of Pennsylvania, Philadelphia, PA 19104, USA and [10]Department of Biochemistry, University of Cambridge, Cambridge CB2 1GA, UK

Received September 26, 2011; Revised and Accepted October 21, 2011

ABSTRACT

GeneDB (http://www.genedb.org) is a genome database for prokaryotic and eukaryotic pathogens and closely related organisms. The resource provides a portal to genome sequence and annotation data, which is primarily generated by the Pathogen Genomics group at the Wellcome Trust Sanger Institute. It combines data from completed and ongoing genome projects with curated annotation, which is readily accessible from a web based resource. The development of the database in recent years has focused on providing database-driven annotation tools and pipelines, as well as catering for increasingly frequent assembly updates. The website has been significantly re-designed to take advantage of current web technologies, and improve usability. The current release stores 41 data sets, of which 17 are manually curated and maintained by biologists, who review and incorporate data from the scientific literature, as well as other sources. GeneDB is primarily a production and annotation database for the genomes of predominantly pathogenic organisms.

INTRODUCTION

GeneDB was initially established with funding from the Wellcome Trust as a curated resource for the kinetoplastid parasites and *Schizosaccharomyces pombe* that were sequenced by the Pathogen Genomics group at the Wellcome Trust Sanger Institute (WTSI) (http://www.sanger.ac.uk). It has now become the home for all pathogen genomes sequenced at the WTSI, as well as several related pathogen genomes from other institutes, and also provides a repository of genome sequences of pathogenic organisms to the wider scientific community. GeneDB provides comprehensive coverage and a systematic approach to pathogen genome annotation, by incorporating and standardizing, data from numerous sources. Currently GeneDB hosts the sequences and associated annotation of 41 organisms, including bacterial, protozoan, helminth and arthropod genomes, making it the largest multi-organism database of curated parasite genomes. The common factor between the genomes is the huge health and economic impact that these organisms have on humans and animals around the world. Malaria is a major cause of human mortality, and is represented on GeneDB by the *Plasmodium falciparum* genome, as well as genomes of several non-human *Plasmodium* species. Chagas disease (American trypanosomiasis), leishmaniasis, schistosomiasis and human African trypanosomiasis

*To whom correspondence should be addressed. Tel: +44 (0)1223 494944; Fax: +44 (0)1223 494919; Email: mb4@sanger.ac.uk
Correspondence may also be addressed to Flora Logan-Klumpler. Tel: +44 (0)1223 495340; Fax: +44 (0)1223 494919; Email: fl2@sanger.ac.uk

(also known as sleeping sickness) are four of the Neglected Tropical Diseases identified by the World Health Organization (http://www.who.int/neglected_diseases/en/), and are represented by genomes in this database. Of the bacterial genomes on GeneDB, 13 are pathogenic to humans. Within the database, there are 34 finished and non-contiguous finished genomes and 7 ongoing sequencing projects (Table 1). Automatic pipelines generate the majority of these genomes, including annotation transfer from closely related, manually annotated species. Several of the eukaryotes have in-depth manual curation, with annotations extracted from the literature and/or provided by users of the website.

Recently, the resource has undergone a complete re-implementation, with many new features added. First, to make curation, and particularly cross-organism curation, easier, the data are stored in a large multi-organism Chado database (2). The use of Chado enables data to be structured in a way that allows rapid, flexible and consistent annotation of any features of a sequence. Second, the website has been modernised to provide a more intuitive and consistent navigation, browsing and searching across all organisms. Currently, the website has an average of 300 unique users per day. Programmatic access to the data has also been extended by means of web services, which are used to drive some of the more dynamic page elements and to provide data upgrades to other linked resources.

GeneDB is a database of annotated genomes or genomes undergoing active annotation. We focus on the rapid release and simple, rapid access, to reference genomes, with a depth of annotation and curation that is not found on other databases [such as ENA-EMBL (3) and GenBank (4)]. Other -omics data is hosted elsewhere, allowing GeneDB to focus on the annotation of reference genomes from the published literature and user comments (see Curation section). This also enables users to explore a genome in detail, particularly in conjunction with the genome-browsing tool, Artemis (see below) (5), with detailed and up to date curation available for every individual gene in a simple to navigate and consistent layout. The simple layout and page design means that data are presented to users extremely rapidly. As part of a collaborative effort with the EuPathDB group of databases (http://www.eupathdb.org) the annotated and curated genomes of the trypanosomatid parasites are regularly sent from GeneDB to TritrypDB (6; http://www.tritrypdb.org) to be integrated with a wide variety of functional genomics data sets such as microarray and proteomics data, which is made available by members of the global research community. The same procedure is used for *Plasmodium* species parasites, which are sent to PlasmoDB (7; http://www.plasmodb.org). In addition, user comments submitted on genes in TriTrypDB and PlasmoDB are automatically forwarded to GeneDB for review and potential integration into the annotation, providing an effective mechanism for reactive curation. A new collaboration with SchistoDB (http://www.schistodb.net) will have the same purpose as those with EuPathDB—to integrate the *Schistosoma mansoni* genome provided by GeneDB with functional genomics data sets from the wider community.

GeneDB has a unique role in producing and improving several key reference genomes, and hence it is a critical part of TriTrypDB, PlasmoDB and more recently, SchistoDB. It is also the platform on which changes, directed from those sites, are implemented. Currently GeneDB is the only database to host multiple flatworm genomes.

DATABASE CONTENT AND CURATION

GeneDB now holds the genome sequences of 9 apicomplexans, 3 of them human pathogens; 12 kinetoplastid protozoans, 7 of them human pathogens; 3 parasitic helminths, all human pathogens; and 16 bacterial species, 13 of them either human pathogens or opportunistic human pathogens. Table 1 shows the current status of the genomes in GeneDB.

Data are currently stored in a PostgreSQL relational database that uses the Chado schema (2), produced by the GMOD consortium. This allows concurrent annotation of multiple organisms in a central data store, and enables curators to make cross-organism annotations.

Several types of annotation are currently performed on genomes once they have been loaded into the database. These can be loosely broken down into automated and manual annotations. Automated annotations resulting from bioinformatics studies are fed into the database via the means of specialised loader programs, wrapped in PERL or BASH scripts. Manual annotations are performed using Artemis, which can connect directly to a Chado database (5). Off-site curators are able to connect our internal database using Artemis, over VPN. Automated annotations include predictions of protein domains [PANTHER (8), PFAM (9), etc.], signal peptides [Signalp 3.0 (10)], GO terms [IprScan (11)], GPI anchors (dGPI; http://129.194.185.165/dgpi), transmembrane helices [tmHMM (12)], non-coding RNA features [RFAMscan (13)], and relationships to orthologous genes in closely related genomes [OrthoMCL (14)]. Results of predictive algorithms are reloaded on GeneDB at regular intervals to keep pace with growing databases.

Much of the information in the database is being comprehensively curated and improved, and is now regularly updated with new assemblies. Manual curation of the database has focused on the trypanosomatids and *Plasmodium* species, and in the past 2 years, more than 11 000 manual curations, based on more than 600 bibliographic references from peer-reviewed international journals have been added (Table 2). GeneDB can be divided into those genomes with proactive curation, including *Trypanosoma brucei 927*, *Trypanosoma cruzi*, *Leishmania major* and *Plasmodium falciparum*, and those with reactive curation. Reactive curation relies heavily on user feedback, whereas proactive curation involves the curators carrying out weekly annotation updates from the published literature and from submitted user comments, but can also involve annotation transfer from the reference genomes to related genomes.

Table 1. GeneDB genomes and curation statuses

		Genome	Curation
Kinetoplastida			
	Bodo saltans		
	Leishmania braziliensis		
	L. infantum		
	L. donovani		
	L. major		
	L. mexicana		
	Trypanosoma brucei 927		
	*T. brucei 427**		
	T. brucei gambiense		
	T. congolense		
	T. cruzi		
	T. vivax		
Apicomplexa			
	Eimeria tenella		
	Plasmodium berghei		
	P. yoelii		
	P. chabaudi		
	P. vivax		
	P. falciparum		
	P. knowlesi		
	Neospora caninum		
Helminth			
	Schistosoma mansoni		
	S. japonicum		
	Echinococcus mulilocularis		
Bacteria			
	Bacteroides fragilis		
	Bordetella bronchiseptica		
	B. parapertussis		
	Burkholderia pseudomallei		
	B. cenocepacia		
	Campylobacter jejuni		
	Chlamydophila abortus		
	Corynebacterium diphtheriae		
	Erwinia carotovora		
	Rhizobium leguminosarum		
	Salmonella typhi		
	S. typhimurium		
	Staphylococcus aureus (MRSA252)		
	S. aureus (MSSA476)		
	S. aureus (TW20)		
	Streptococcus pnuemoniae		
	S. pyogenes		
	Streptomyces coelicor		

Project Type
Improved high-quality draft
Noncontiguous finished
Finished

Curation
Reactive curation
Proactive curation
No curation

Project type: Yellow indicates improved high-quality draft; cyan indicates non-contiguous finished; green indicates finished. Curation: blue indicates reactive curation; hatched blue indicates proactive curation; No curation white.
Project type is based on categories defined in (1).
[a]Not a whole-genome project.

Table 2. Snapshot of annotation and curation—statistics from a 12-month period between August 2010 and August 2011

Annotation event type	*Leishmania major*[a]	*Plasmodium falciparum*	*Trypanosoma brucei 927*	*Trypanosoma cruzi*
Assigned or updated Product	304	293	241	391
Updated GO term	751	121	1718	1163
Phenotype curation term added	165	–	6791	315
Linked to publication (PMID)	496	244	1456	693
User Comment added	252[b]	110	312[b]	519[b]
All unique genes with new functional annotations	1220	839	8750	1869
All unique genes with new structural annotations	13	291	41	–

Annotation event types are shown for each of the four reference genomes in GeneDB.
[a]*L. major* represents ~50% of all *Leishmania* species curation activity.
[b]User comment entered at Tritrypdb.org.

The database also contains a large collection of fully finished bacterial genomes. These are predominantly browsable snapshots of published genomes and reference material for functional or comparative studies. Curation activities currently focus on the preparation of new sequence releases, with updated gene structures and functional annotations of gene products and mutant phenotypes, as well as other experimental phenotype data. Publications that describe new gene features in the organisms included in the GeneDB data are annotated to the database, with an emphasis on the reference genomes already described. Much of the annotation is gleaned from published literature; although some is taken from user comments (see Interactions with other resources section).

Curation and annotation of the kinetoplastids and *Plasmodium* species in GeneDB is a collaborative effort, with teams of scientists and bioinformaticians at three institutions working closely with each other. Genomic annotations are prepared via rapid information and knowledge exchange between teams of literature annotators and data curators. Teams follow an open information management infrastructure using web-based applications: Zotero (http://www.zotero.org; a bibliographic and research information management system), JIRA (traditionally used for software development bug tracking, but used here for tracking user comments and curation tasks), and Google groups. Dissemination of Kinetoplastida genome knowledge in particular is accelerated through the innovative use of these web-based tools as an open information-sharing platform.

ANNOTATING WITH ARTEMIS

Curators from all three sites access the database using Artemis (5) and modifications are individually logged and time stamped in the database. Annotations include literature and other database cross-references, Gene Ontology (GO) (15) terms inferred from the literature and user comments, and phenotype curations, using a semi-controlled vocabulary. Evidence codes are used according to GO protocols. To maximize the potential of Artemis as a tool for genome annotation, it is able to read and write directly to the Chado relational database schema. In addition, the Gene Builder tool provides structured forms and tables to edit coordinates of gene models and edit functional annotation, based on standard ontologies, GeneDB unique controlled vocabularies (CVs), and the option to use free text when necessary (16). There are four annotation sections in the Gene Builder that enable curators to capture descriptions of gene products, related phenotypes, GO terms and orthologues/paralogues. This provides the end users with a comprehensive overview of what a gene does, where it is found, and which genes it is most like in related organisms.

CONTROLLING VOCABULARIES WITH THE RATIONALISER

Chado uses CVs to ensure that annotations conform to agreed standards and disambiguate the way the data sets are queried and interpreted by different research communities. Able to represent ontologies, by the means of CV term relationships, it can make use of any of the ontologies for annotating biological data (Open Biological Ontologies; http://www.obofoundry.org). However, in some cases these do not provide the specific terms needed, such as gene product names and descriptions, giving rise to in-house CVs (populated from free-text lists maintained by the curators). In order to avoid redundancy, inconsistencies need to be identified and corrected (e.g. some product annotations can read ‘kinase, putative’, while others will read ‘putative kinase’).

The Rationaliser (to be described in detail elsewhere, http://rationaliser.sourceforge.net/) is an interactive tool that enables curators to easily refine these CVs within a set of organisms of interest (Figure 1). For instance, a curator may choose to rationalise product names across annotations in all the *Plasmodium* species. Once connected to the database, the Rationaliser presents the user with two lists: a list of terms from a chosen CV that are used across the specified set of organisms, and a list of all the terms in the entire CV. Curators search through these lists and, for each inconsistent term, choose a replacement from the other terms in the CV. The Rationaliser then updates the CV and any annotations that had made use of the incorrect term. If the right term does not exist in the CV yet, the user can also create a new term. Terms are colour-coded by evidence code, to help curators establish

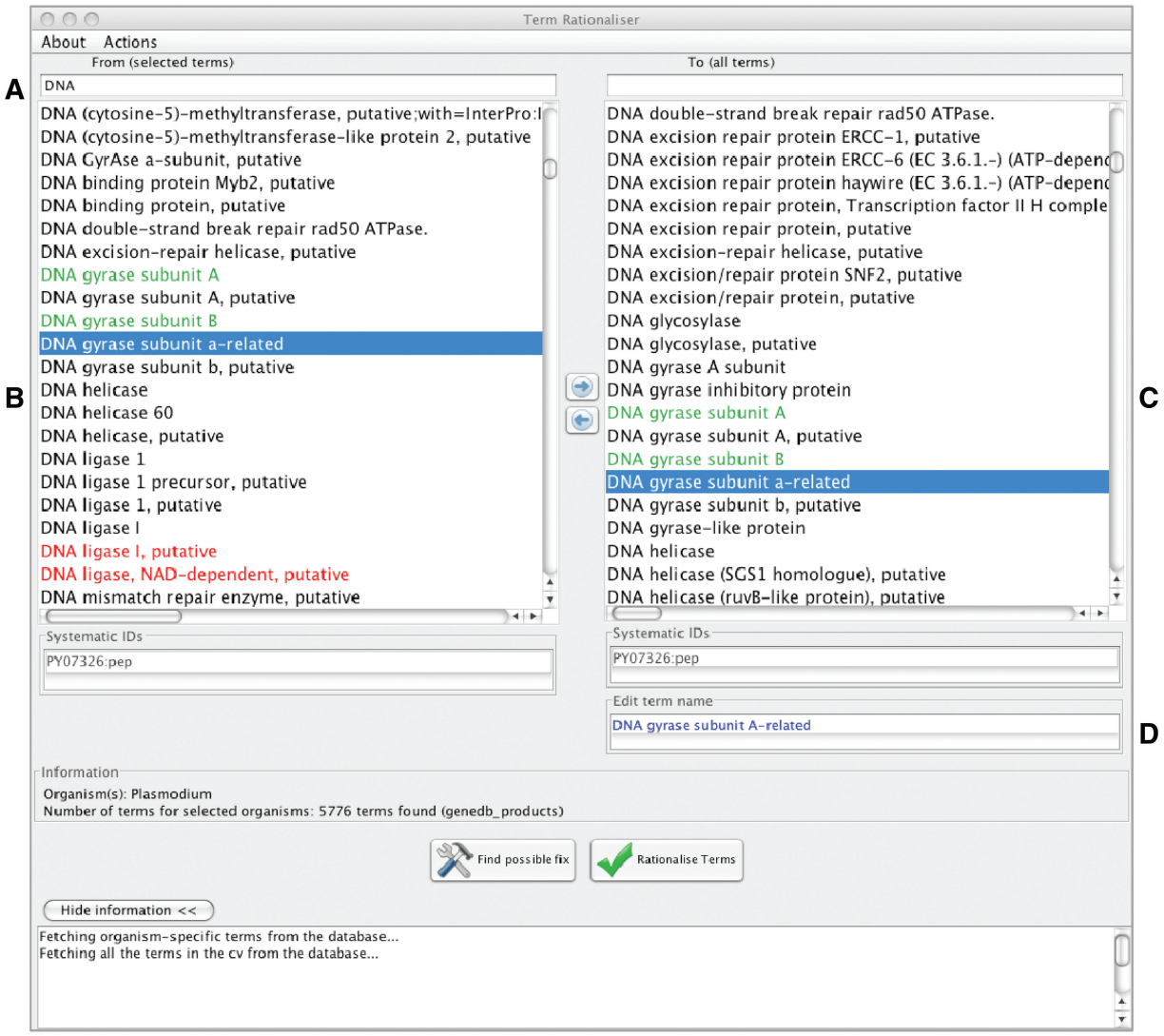

Figure 1. The Rationaliser tool is used to remove inconsistencies from curated data. (**A**) Search boxes to locate terms. (**B**) List of colour-coded CV terms used within the annotations of selected organisms. (**C**) List of all available terms in the CV (also colour-coded). (**D**) Text box to insert new terms if the correct term does not exist.

their credibility. The Rationaliser therefore gradually, and iteratively, helps improve in-house CVs. Organism-specific curators in the Pathogens group at the WTSI and Seattle Biomed use the Rationaliser to fix inconsistencies in gene product descriptions in GeneDB.

REMAPPING WITH RATT

Second generation sequencing is facilitating the production of new assemblies of organisms, meaning frequent sequence updates. This poses a problem of having to relocate genes. Likewise, first versions of strains similar to existing reference genomes now appear more frequently too, with similar chromosomes, and similar genes. The Rapid Annotation Transfer Tool (RATT, 17) takes advantage of NUCMER matches to transfer gene model predictions and product annotations between genomes. It currently runs off EMBL files so to run this, the data is exported out of the database.

In the simplest case, for assemblies of new strains, the annotations are transferred across to the new genome, and the resulting EMBL file is loaded using a program. It is harder to deal with updates of existing genomes, because in these cases genes move, some new ones are created, and some are made obsolete. RATT is used to recalculate the coordinates of the pre-existing genes, and the genes are moved in the database accordingly. New genes are then loaded, and old ones made obsolete.

This illustrates a further advantage to maintaining a large database of genomes for which the initial annotation was manually performed. That is the potential to transfer refined gene model predictions from finished genomes to draft genomes. This has been facilitated recently in the recent annotation of the *Leishmania donovani* genome (18); 8252 of the 8395 *L. infantum* genes predictions

could be transferred to the unannotated *L. donovani* genome by RATT.

INTERACTIONS WITH OTHER DATABASES

GeneDB has a close relationship to other databases, those of EuPathDB in particular. It provides the underlying genome and annotation data on which PlasmoDB, TriTrypDB and some of ToxoDB are built. GeneDB has a far shorter release cycle than those of the EuPathDB databases, and as such always contains the most up-to-date version of a genome. This is in part because GeneDB generates and hosts only genomic data (as opposed to the more complex collections of data sets hosted by EuPathDB), with the focus on keeping these essential data sets accurate and current. Large-scale data releases are made to EuPathDB by means of GFF3 files, which are bulk-exported using Artemis' command line writedb_entry tool. In between releases, EuPathDB call our web-services daily to determine what new annotations have taken place on GeneDB, and use this information to flag up the presence of annotation changes on their gene pages, linking back to our Gene pages.

In conjunction with EuPathDB, GeneDB curators have encouraged members of the scientific community to submit user comments to enhance the curation of the database. Comments made by scientific community members not only appear directly on the EuPathDB gene page, but are also used, when relevant, to improve annotations directly in the database. This collaboration with TriTrypDB and PlasmoDB enables the submission and management of user comments to be fed back to the GeneDB annotation team. To date, over 4500 comments have been made on genes in *Plasmodium* species and across the kinetoplastids, and 1618 of these have been used to update gene annotations. This community aspect of curation will be developed further at GeneDB, in conjunction with PomBase (19; http://www.pombase.org).

GeneDB has also provided Ensembl with up-to-date genome information for several species, e.g. *L. major* and *P. falciparum*, and will continue to build on this relationship with the sharing of further genomes. We also supply the genome and annotation data for SchistoDB.

We also provide trackback links to numerous other biological databases. For *Plasmodium* a number of databases can be accessed via appropriate gene pages, for example, RMGM, a database of genetically modified malaria parasites and MPMP, a database for malaria parasite metabolic pathways. It is also possible to access databases such as KEGG and UniProt through links on gene pages on GeneDB.

PROGRAMMATIC DATA ACCESS AND WEB SERVICES

There are several methods through which one can access data on GeneDB. The public read-only snapshot database can be accessed via any PostgreSQL client, and therefore most programming languages, but requires knowledge of the Chado schema. If ones language of choice runs on the JVM, it is entirely feasible to user either the GeneDB or Crawl (Chromosomal Resource Annotation Web-Service Layer) projects (https://github.com/sanger-pathogens/) as libraries to get access to the data.

Crawl (to be described in detail elsewhere) is a dedicated web services application, exposing the data on GeneDB via both REST-like resources and SOAP endpoints. Its REST-like resources return data as XML and JSON. The XML is used by collaborators at EuPathDB to let their users know of updated annotations on GeneDB. The JSON is used by graphical widgets on GeneDB that are written in Javascript, and is very easy to consume in any programming language. Because of their self-describing nature, the SOAP endpoints are better suited for use in web-service-aware workflow-based analysis environments, such as Pipeline Pilot or Taverna. Crawl services provide a comprehensive API to the rich information stored in GeneDB, and are accessible from (http://www.genedb.org/services/).

NEW WEBSITE DESIGN AND FEATURES

The new GeneDB website is more graphical than previous versions, with images representing the general biological areas we focus on. Beneath each icon is a clickable dropdown list of organisms, which enables users to navigate to the homepage of their organism of interest (Figure 2). These organism homepages are modelled on the organism homepages in old GeneDB (20,21), with links to search tools, other databases of interest and entry points into the genomes. There is also a link to an information page, which provides a background to the organism and its genome project, along with the publication details for researchers to reference. On the website, the organisms are arranged into a hierarchy, which is not a strict phylogenetic relationship, but reflects the areas on which users are most focused, and across which they normally wish to carry out comparative analyses. This hierarchy is reflected both in the navigation of the website, and in the searching. Both the queries and the browsable lists can be run against any node in the hierarchy to allow either a focus on one organism, or across a related family.

New features on the organism homepage include annotation statistics, which allows users to ascertain when the most recent annotations were carried out on this genome, and click through to the most up to date gene pages (Figure 3). Chromosome and contig maps were very popular features on previous versions of GeneDB. New versions of these have been developed, with a fresher, data-driven design. They therefore are be updated when new annotations are made, which is an improvement on the previous, static, versions. These maps are very useful entry points to the genome, allowing users to scroll through the chromosomes, see genes in the context of their neighbours, and the broader chromosome structure, as well as clicking through to genes of interest.

Various searches are available, allowing the user to find genes of interest based on their annotation characteristics, each with options to restrict on an organism group.

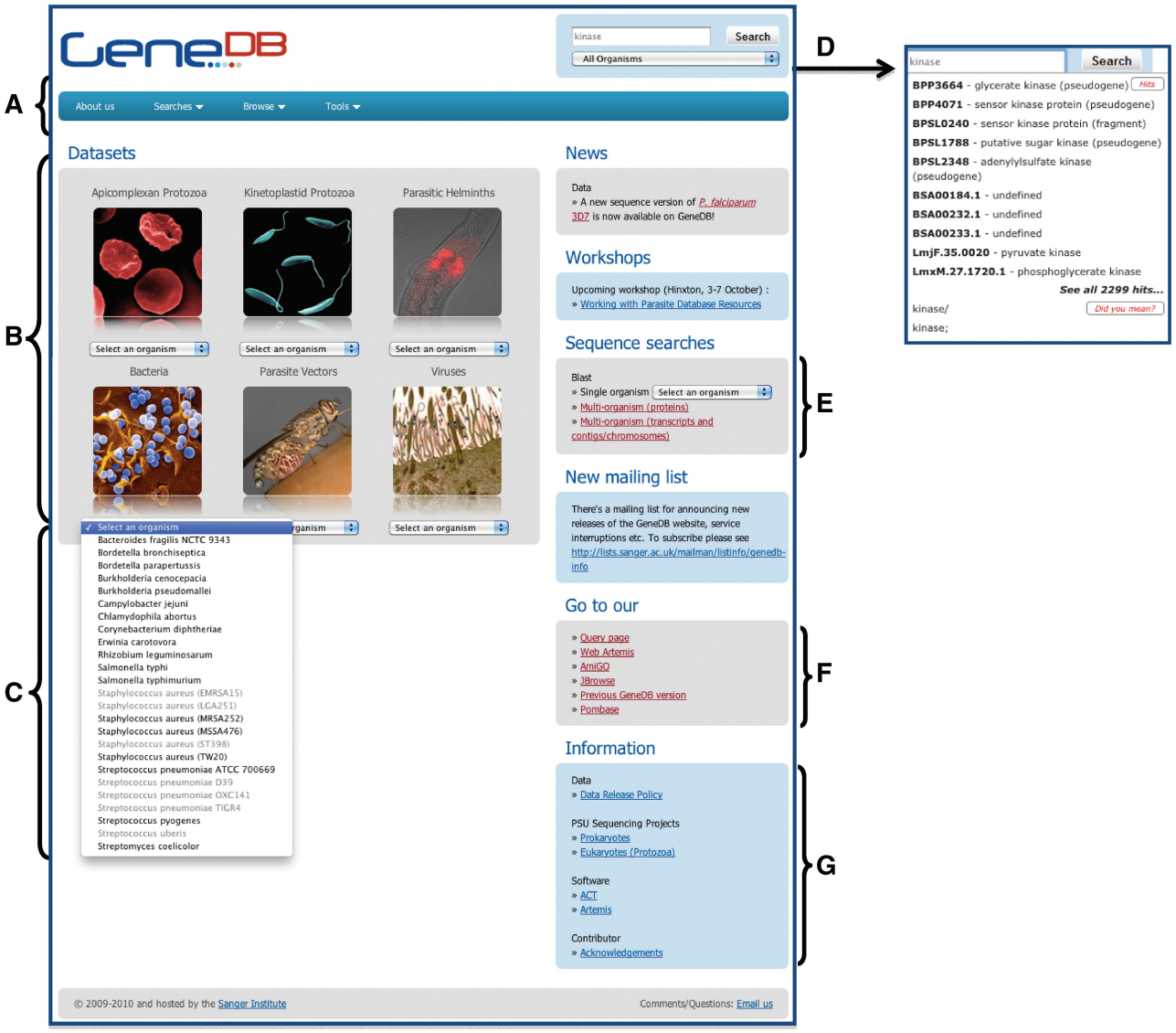

Figure 2. Screen shot of GeneDB homepage with an example of an entry point into an individual organism homepage. (**A**) Links to available tools. (**B**) Available data sets. (**C**) Clicking on 'Select an organism' opens a drop-down box with available bacterial genomes. (**D**) Quick search option. (**E**) Blast search options. (**F**) Links to available tools. (**G**) Links to ongoing projects and data release policy.

The results can be browsed, linking to gene pages, and also downloaded. A new feature is the quick search, which occupies a prominent position on the top-right of every page, and auto-suggests both direct matches and fuzzy matches to gene names and products as one types. This enables researchers to quickly get to their gene of interest.

The gene page displays details of a chromosomal feature, usually a gene (Figure 4). For alternatively spliced genes, there exists a gene page for each spliced variant. Also non-gene elements can be annotated and therefore have pages e.g. ncRNAs, gaps or centromeres. The gene page sections are broadly similar to the previous version of the page, with minor changes based on community feedback. The context map at the top of each gene page has been completely rewritten, using Web-Artemis. Where previously the map was static, Web-Artemis can be navigated across the chromosome, zooming in and out of areas of interest, providing a simple, gene-centric, view of the contig. An overview above the map shows the relative position of the view on the gene page, which can be dragged. More detailed views of the structure and annotation are available through Artemis and Web-Artemis buttons. Another button takes one to the corresponding sequence for the feature on a separate page, which includes the genomic sequence, and protein sequence where applicable, along with quick links to analyse them again using BLAST.

WEB-ARTEMIS

Web-Artemis (to be described in detail elsewhere) is a genome graphical viewer, that we use embedded in the gene pages to help navigate chromosomes, as well as on

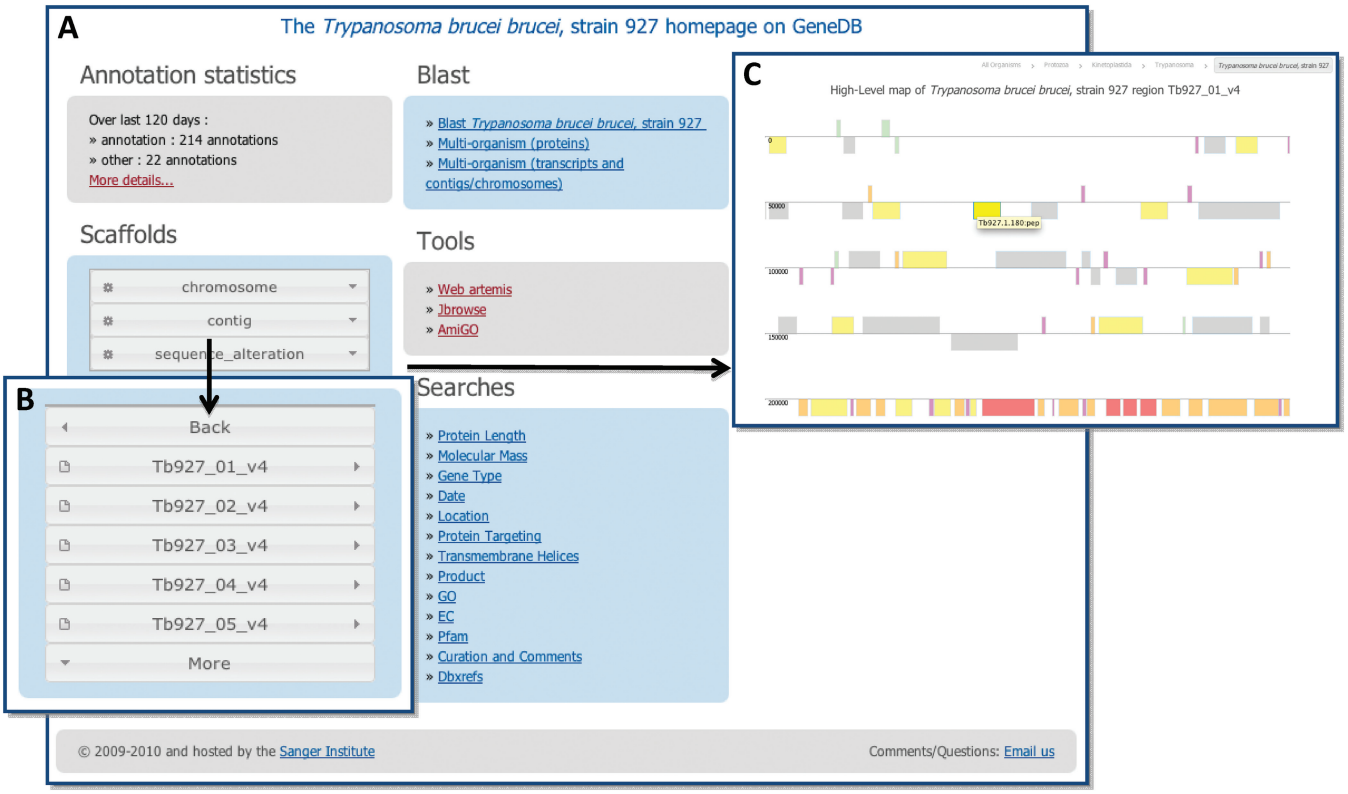

Figure 3. Genes can be accessed by browsing clickable whole chromosome or contig maps. (**A**) Screen shot of an organism home page showing annotation statistics, blast tools and search tools. (**B**) List of all chromosomes. (**C**) Shown is part of a chromosome map. A mouse over shows the systematic IDs of the genes, a click on the gene of interest opens the gene page.

standalone pages, where its alignment (BAM) and variant call format (VCF) display functions will allow us to overlay next-generation sequencing efforts on top of the annotations. It is a pure Javascript application that makes AJAX requests to Crawl services for getting its data. It aims to provide the user with a view similar to (but simpler than) the view that Artemis provides, but rendered solely in the browser.

ARTEMIS

A link to Artemis (the standalone application) is also present on every gene page, allowing a very detailed view of the gene and its surrounding features to be accessed, analysed and downloaded using one of its many data export functions for further in depth analysis.

OTHER TOOLS

Installations of BLAST, Amigo and JBrowse are maintained, populated with data from our database.

THE AFRICAN TRYPANOSOME CELL-SURFACE PHYLOME

The African Trypanosome Cell-Surface Phylome is a collection of Bayesian and Maximum Likelihood phylogenies created for gene families with predicted expression in or on the cell-surface, and present in the genomes of *T. brucei* TREU927, *T. congolense* IL3000 and *T. vivax* Y486.

(http://www.genedb.org/Page/trypanosoma_surface_phylome).

Cell-surface expression has been predicted for genes encoding signal peptides, glycosylphosphatidylinositol (GPI) anchors or transmembrane helices. Putative cell-surface genes have been placed into 81 families by searching trypanosome genomes using wuBLAST, manually curating sequence alignments to ensure genuine homology, and removing genes known to function away from the cell-surface, e.g. across organelle membranes. The phylome shows that while some genes families such as ABC transporters (Fam56), major facilitator superfamily transporters (Fam58) and metal ion transporters (Fam55) have evolved with almost no changes in gene repertoire, species-specific derivations of gene families familiar for their roles at the host-parasite interface such as major surface proteases (Fam46), *trans*-sialidases (Fam47), cysteine peptidases (Fam67), and the variant surface glycoproteins [VSG (22)], have been considerable. Finally, the phylome identifies species-specific cell-surface gene families. In *T. brucei* many of these genes are Expression Site-Associated Genes [ESAGs (23)] or otherwise GPI-anchored hypothetical proteins. In *T. vivax* there are 23 species-specific families; with the exception of *VSG*-like genes (Fam23-26), these families have no homologous sequences in other species. Taken

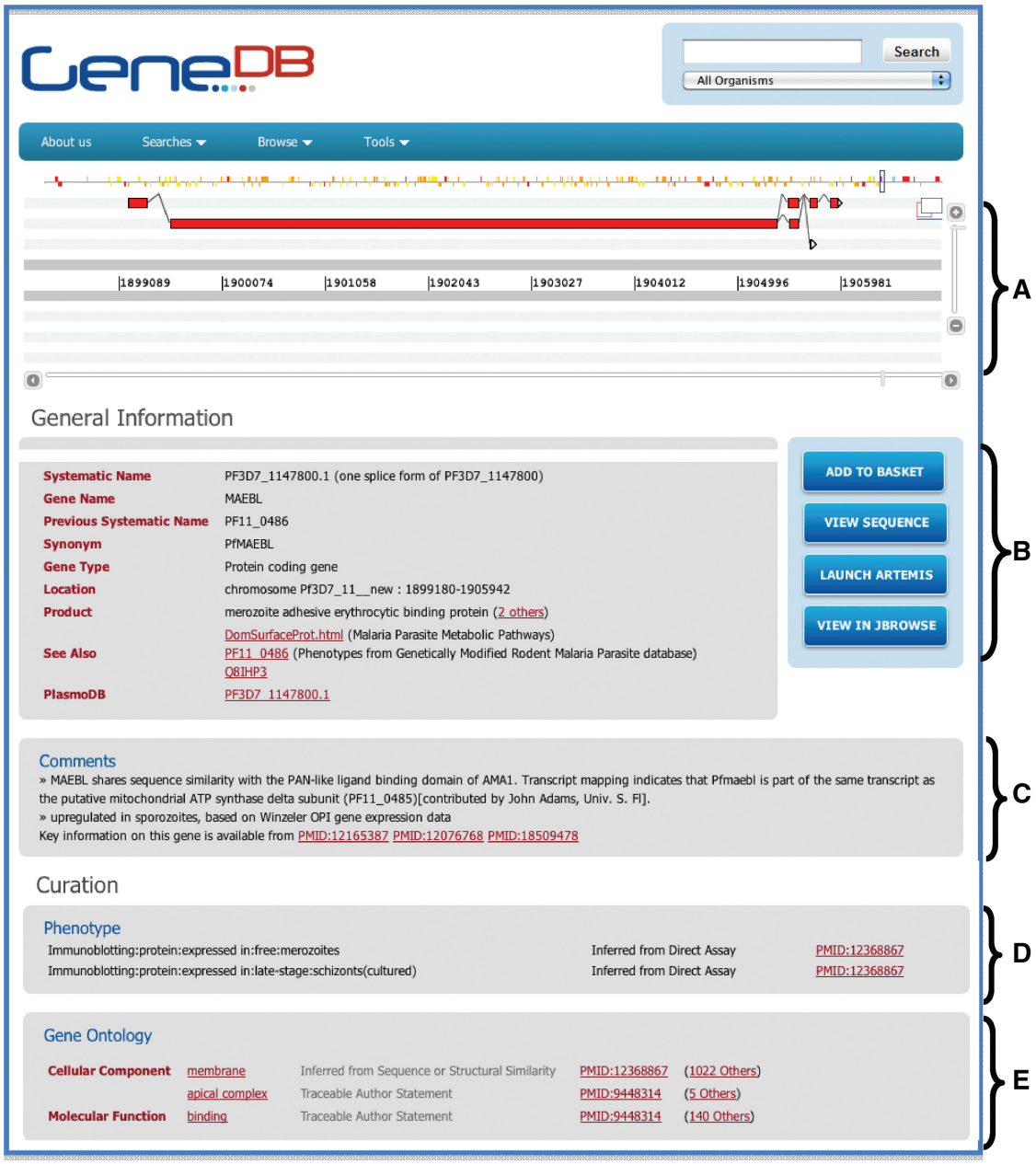

Figure 4. Screen shot of GeneDB gene page. (**A**) Graphical display of the gene in web-artemis. The chromosomal location of the gene is shown on the top. (**B**) Basic gene information. (**C**) Additional information, including relevant literature. (**D**) Phenotype information. (**E**) Gene Ontology terms.

together, the phylome emphasizes the substantial genomic innovation that has affected trypanosome cell surface architecture, which should be reflected in phenotypic variation in host-parasite interactions.

FUTURE DIRECTIONS

With the expanding use of new sequencing technologies, we are expecting to load an accelerating number of organisms. Here the focus has shifted away from the generation of new reference genomes, to the mapping of variant data to existing references. Currently short-read alignments can be viewed using Artemis BAMview (5), and more recently pileup format files, converted to VCF, can now be viewed with Artemis VCFview. This has opened the potential for viewing SNP calls for multiple related species/strains in the same window, aligned against a reference genome. New interfaces will be developed for querying and displaying these data. We also intend to further develop our querying interface. This will provide the means to manipulate result sets after creation, for example, editing or merging them. We will also provide new download options, based, as the current set are, on user feedback.

Second generation sequencing technology has also allowed for deep re-sequencing of existing reference genomes. Re-sequencing of pathogen genomes has

recently taken on an added importance, as with the growing number of large-scale diversity studies taking place, it is necessary to discriminate between SNP predictions owing to polymorphisms between strains/isolates, and those resulting from mis-called bases in the reference genome. Recent re-sequencing of *Leishmania* genomes using Illumina technology has increased the existing depth of these genomes to over 50×, and up to 100× in *L. braziliensis* and *L. mexicana* (24). Using the iterative mapping/SNP prediction algorithm, Iterative Correction of Nucleotide Sequences (ICORN) (25), thousands of mis-called bases and indels have been corrected in each of the *Leishmania* genomes. These improvements in sequence quality have resulted in positive changes to gene model predictions. In *L. infantum* as an example, 98 gene models previously containing internal frameshifts were converted to intact open reading frames following ICORN improvement.

As the volume of publications concerning pathogenic organisms continues to increase, complete manual curation remains a challenge with current methods. Building on the user comments received through the EuPathDB collaboration, we will expand and formalise this method of community curation with a custom built tool. This will enable users to enter annotations from their own research using ontologies, as well as information about gene structures and other sequence updates. The curation of phenotype data is currently reliant on a semi-controlled vocabulary, and this lexicon is being used to develop a phenotype descriptor ontology specifically for use in GeneDB curation. This Neglected Tropical Diseases Phenotype Ontology (NTDPO) will be used in combination with the Ontology of Parasite Lifecycles (OPL; http://bioportal.bioontology.org/ontologies/39544), and GO, to generate a wholly consistent vocabulary to describe experimental phenotype data in GeneDB organisms.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the contribution of numerous members of the parasitology research community, in the form of advice, suggestions and/or data.

FUNDING

Wellcome Trust (grants WT 043565 to M.B., M.C., D.R. and D.F.S.); (WT 085775/Z/08/Z to the Wellcome Trust Sanger Institute). Funding for open access charge: Wellcome Trust Sanger Institute.

Conflict of interest statement. None declared.

REFERENCES

1. Chain,P.S., Grafham,D.V., Fulton,R.S., Fitzgerald,M.G., Hostetler,J., Muzny,D., Ali,J., Birren,B., Bruce,D.C., Buhay,C. *et al*. (2009) Genomics. Genome project standards in a new era of sequencing. *Science*, **326**, 236–237.
2. Zhou,P., Emmert,D. and Zhang,P. (2006) Using chado to store genome annotation data. *Curr. Protocols Bioinformatics*, **9**, 9.6.1–9.6.28.
3. Leinonen,R., Akhtar,R., Birney,E., Bower,L., Cerdeno-Tarraga,A., Cheng,Y., Cleland,I., Faruque,N., Goodgame,N., Gibson,R. *et al*. (2011) The European nucleotide archive. *Nucleic Acids Res*., **39**, D28–D31.
4. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2011) GenBank. *Nucleic Acids Res*., **39**, D32–D37.
5. Carver,T., Berriman,M., Tivey,A., Patel,C., Böhme,U., Barrell,B.G., Parkhill,J. and Rajandream,M.A. (2008) Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics*, **24**, 2672–2676.
6. Aslett,M., Aurrecoechea,C., Berriman,M., Brestelli,J., Brunk,B.P., Carrington,M., Depledge,D.P., Fischer,S., Gajria,B., Gao,X. *et al*. (2010) TriTrypDB: a functional genomic resource for the *Trypanosomatidae*. *Nucleic Acids Res*., **38**, D457–D462.
7. Aurrecoechea,C., Brestelli,J., Brunk,B.P., Dommer,J., Fischer,S., Gajria,B., Gao,X., Gingle,A., Grant,G., Harb,O.S. *et al*. (2009) PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Res*., **37**, D539–D543.
8. Thomas,P.D., Kejariwal,A., Campbell,M.J., Mi,H., Diemer,K., Guo,N., Ladunga,I., Ulitsky-Lazareva,B., Muruganujan,A., Rabkin,S. *et al*. (2003) PANTHER: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic Acids Res*., **31**, 334–341.
9. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al*. (2010) The Pfam protein families database. *Nucleic Acids Res*., **38**, D211–D222.
10. Bendtsen,J.D., Nielsen,H., von Heijne,G. and Brunak,S. (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol*., **340**, 783–795.
11. Quevillon,E., Silventoinen,V., Pillai,S., Harte,N., Mulder,N., Apweiler,R. and Lopez,R. (2005) InterProScan: protein domains identifier. *Nucleic Acids Res*., **33**, W116–W120.
12. Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L.L. (2001) Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J. Mol. Biol*., **305**, 567–580.
13. Gardner,P.P., Daub,J., Tate,J., Moore,B.L., Osuch,I.H., Griffiths-Jones,S., Finn,R.D., Nawrocki,E.P., Kolbe,D.L., Eddy,S.R. *et al*. (2011) Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Res*., **39**, D141–D145.
14. Chen,F., Mackey,A.J., Stoeckert,C.J. Jr and Roos,D.S. (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res*., **34**, D363–D368.
15. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al*. (2000) The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nat. Genet*., **25**, 25–29.
16. Rutherford,K., Parkhill,J., Crook,J., Horsnell,T., Rice,P., Rajandream,M.A. and Barrell,B. (2000) Artemis: sequence visualization and annotation. *Bioinformatics*, **16**, 944–945.
17. Otto,T.D., Dillon,G.P., Degrave,W.S. and Berriman,M. (2011) RATT: rapid annotation transfer tool. *Nucleic Acids Res*., **39**, e57.
18. Downing,T., Imamura,H., Decuypere,S., Clark,T.G., Coombs,G.H., Cotton,J.A., Hilley,J.D., de Doncker,S., Maes,I., Mottram,J.C. *et al*. (2011) Whole genome sequencing of multiple *Leishmania donovani* clinical isolates provides insights into the evolution and mechanisms of drug resistance. *Genome Res*.
19. Wood,V., Harris,M.A., McDowall,M.D., Rutherford,K., Vaughan,B., Staines,D.M., Lock,A., Aslett,M., Bähler,J., Kersey,P. *et al*. (2012) PomBase: a comprehensive online resource for fission yeast. *Nucleic Acids Res*., **40**, D695–D699.
20. Aslett,M., Mooney,P., Adlem,E., Berriman,M., Berry,A., Hertz-Fowler,C., Ivens,A.C., Kerhornou,A., Parkhill,J., Peacock,C.S. *et al*. (2005) Integration of tools and resources for display and analysis of genomic data for protozoan parasites. *Int. J. Parasitol*., **35**, 481–493.
21. Hertz-Fowler,C., Peacock,C.S., Wood,V., Aslett,M., Kerhornou,A., Mooney,P., Tivey,A., Berriman,M., Hall,N.,

Rutherford,K. *et al*. (2004) GeneDB: a resource for prokaryotic and eukaryotic organisms. *Nucleic Acids Res*., **32**, D339–D343.
22. Borst,P. and Cross,G.A. (1982) Molecular basis for trypanosome antigenic variation. *Cell*, **29**, 291–303.
23. McCulloch,R. and Horn,D. (2009) What has DNA sequencing revealed about the VSG expression sites of African trypanosomes? *Trends Parasitol*., **8**, 359–363.
24. Rogers,M.B., Hilley,J.D., Dickens,N.J., Wilkes,J., Bates,P.A., Depledge,D.P., Harris,D., Her,Y., Herzyk,P., Imamura,H. *et al*. (2011) In press. Chromosome and gene copy number variation allow major structural change between species and strains of Leishmania. *Genome Res*.
25. Otto,T.D., Sanders,M., Berriman,M. and Newbold,C. (2010) Iterative correction of reference nucleotides (iCORN) using second generation sequencing technology. *Bioinformatics*, **26**, 1704–1707.